A phenomenon of contrasting explicit learning effects in verbal statistical learning

Lu Wang[a], Tianlin Wang[b], Wenbo, Yu[a], Dandan Liang[a,c]*

[a] School of Chinese Language and Culture, Nanjing Normal University, Jiangsu Province, P.R. China

[b] University at Albany, State University of New York, New York, USA

[c] Interdisciplinary Research Centre for Linguistic Science, University of Science and Technology of China, Anhui Province, P.R. China

* Corresponding author

Telephone: (+86) 13815866738

E-mail: ldd233@sina.com

Abstract

When accessing the learning effect in statistical learning, participants are requested to distinguish target words from partwords or nonwords in a two-alternative forced-choice task. However, this task did not answer how individuals represent target words and foils, and thus may not be sufficient in providing an independent learning effect on the items. The current study examined the explicit learning effect for each word type with a familiarity rating task. Participants were randomly assigned to learn a continuous artificial speech in one of three conditions: a baseline, short-exposure time, or long-exposure time condition. The ratings scores and correlations across three types of words between the baseline condition and the other two learning conditions were examined. Results revealed a significant contrasting learning effect: familiarity ratings for target words were significantly higher than baseline, whereas foils' ratings were significantly lower, reflecting the explicit knowledge during the exposure phase and the metacognition during the testing process. Furthermore, the distribution of target words' rating scores tends to be more centralized in the long-exposure time condition, suggesting a new type of SL effect. This study is the first to explore explicit learning effects across word types, and provides insight regarding how to measure SL more exactly.

Keywords: statistical learning, learning effect, explicit knowledge, familiarity rating task, two-alternative-forced task

1. Introduction

Statistical learning (SL) refers to the capacity to recognize statistical patterns and identify cognitive units, which are fundamental aspects of cognition (Saffran et al., 1996). In this context, verbal SL has traditionally been recognized as playing a key role in segmenting words from continuous speech. While substantial research has linked SL to the development of language skills and reading abilities (e.g., Shoaib et al., 2018; von Koss Torkildsen et al., 2019; Qi et al., 2019; Frost et al., 2019; Isbilen et al., 2022; Lukács et al., 2023), some fundamental yet controversial issues remain, drawing significant attention from researchers: Is statistical learning an implicit or explicit cognitive mechanism? Is there a better measurement method than the two-alternative-forced-choice (2AFC) task for adults? Does the length of exposure affect learning effects?

1.1 Is statistical learning an implicit or explicit mechanism?

The most prevalent paradigm in SL research is the "learning-testing" paradigm. In learning/exposure phase, participants are implicitly required to learn an artificial language with statistical regularities. For example, in the study by Mirman et al. (2008), participants were asked to listen to a "made-up language" and "be ready to answer questions about the language", while in the study by Wang and Saffran (2014), they were told to "pay as much attention to the language pattern as possible". In the testing phase, researchers have typically used a 2AFC task for both child and adult groups, where participants are told to choose the one that sounds more familiar. Since neither the instructions in the exposure phase nor those in the test phase made it clear that there were units to be segmented from the continuous speech, the entire design is considered to be tapping to an implicit learning mechanism.

Recently, the view that statistical learning is not a purely implicit learning mechanism has become increasingly popular. For example, Batterink et al. (2015a) examined the explicit and implicit groups' learning effects by using a target-detection task, where participants were first exposed to a continuous stream of repeating nonsense words and were then instructed to detect a specific syllable within a continuous speech stream. The results showed that explicit group of participants, who received a supplementary explicit training on the nonsense words, responded faster and more accurately when the target syllable was in the third position than implicit group. In another study, also from Batterink et al. (2015b), participants in explicit group were told to figure out the words' boundaries in a nonsense language and test their knowledge, whereas participants in implicit group were simply instructed to listen to the auditory stimuli. Despite significant learning effects observed in both recognition and target-detection tasks among participants, these effects failed to correlate with each other, thereby underscoring the complexity of SL effects.

The confidence rating task, where participants are asked to indicate how sure they are in their response on a 4-point scale or 7-point scale, provides more evidence that explicit knowledge is involved in SL mechanism. One study explored how adults abstract different types of statistic regularities simultaneously; and results revealed that participants were more aware of their knowledge of the non-adjacent statistics regularities than adjacent regularities (Romberg & Saffran, 2013). Ordin and Polyanskaya (2021) designed a 2AFC task with three types of pairs: (1) target words against phantoms (triples that are not embedded as whole units in the familiarization input but are statistically congruent with words), (2) target words against nonwords (created by utilizing non-adjacent syllables within the artificial language), and (3) phantoms against nonwords. Then, they estimated participants' metacognition scores based on their 2AFC task

performance and confidence rating scores. It is not surprising that participants' discrimination between words and nonwords are similar to that of phantoms and nonwords, while metacognitive sensitivity is higher in trials in which target words are paired with nonwords than trials in which phantoms are paired with nonwords. This indicates that participants could identify the target words consciously. These empirical studies provide direct and compelling evidence that learning effects in SL indeed encompass explicit knowledge.

1.2 Is there a better measurement method than the 2AFC task for SL effects?

Related to the discussion about the learning mechanism of SL, many researchers have claimed that different test tasks reflect different learning outcomes. As a rule, performance in recognition measures, like 2AFC and statistically induced chunking recall, requires the actual extraction and memorization of word units in learning phase, that is, explicit statistical knowledge. In contrast, outcomes from syllable target-detection tasks reflect the extent to which participants are sensitive to the statistical information, that is, the implicit statistical knowledge. A growing body of studies has confirmed the low and non-significant correlation between 2AFC and target-detection tasks (e.g., Batterink, 2017), but a high and significant relationship between 2AFC and statistically induced chunking recall in both adult and child groups (Isbilen et al., 2020; Isbilen et al., 2022).

However, comparatively little attention has been given to the drawbacks of the 2AFC task. When utilizing the 2AFC task, three distinct word types are involved: target words, partwords (formed by combining consecutive syllables from two target words), and nonwords (formed by syllabic sequences that never occurred during the exposure phase). As participants engage in the 2AFC task, they are confronted with pairs consisting of target words paired with either partwords or nonwords. Consequently, the scores derived from this task reflect a compound score, where

the learning outcomes of both the target words and the foils (partwords and nonwords) can potentially contribute. This highlights that, while the 2AFC task is solely used to detect SL learning effects, it only provides information about the ability to differentiate between target words and foils, lacking the ability to independently differentiate results based on item type. In addition, each artificial languages in a SL study is often created with four to six target words. Thus, to increase the number of forced-choice trials, each target word has to be repeated to pair with multiple foils, sometimes as many as six times, as in Batterink et al. (2015a), introducing a secondary learning effect produced in test phase (Siegelman et al., 2017). Especially when participants must distinguish between a word and a partword that differs by only a single syllable, repeated exposure to the same target word may increase participants' confidence in making choice, thereby exaggerating the learning effect observed during the exposure phase. In sum, when measuring explicit knowledge of SL, the 2AFC task not only has limited ability to assess the learning effects of individual word types within the SL task but also comes with inherent limitations.

The familiarity rating task emerges as a viable alternative to the conventional 2AFC task. This task has already been explored in prior studies. For instance, Batterink and Paller (2017) identified a linear decline in participants' learning across three distinct word types, reflecting participants' ability to abstract statistical regularities. In addition, Erickson et al. (2016) discovered that participants' performance in both the 2AFC task and the familiarity rating task exhibited correlation in specific versions of artificial languages. Crucially, unlike the 2AFC task, the familiarity rating task allows participants to independently form explicit memory representations for different word types, enabling them to assess their familiarity with each individual item separately. Finally, the familiarity rating task requires each word to be presented

only once, thereby avoiding secondary learning effects. This study adopts this approach to evaluate independent explicit knowledge of each word type.

To accurately assess the independent learning effects, a direct comparison of score differences among the three types of words is insufficient. Instead, a novel approach that can more effectively isolate and analyze these effects should be devised and considered. This study introduces a baseline condition that is influenced by the work of Toro et al. (2011). In this baseline condition, the artificial language was composed of nonsensical syllables from the same pool used in the experimental condition. Because none of the three types of words appeared in the exposure phase of this condition, memory representations for these items are anticipated to remain at baseline levels, allowing a clean comparison of learning effects for each of the three types of words. Another rationale for incorporating a baseline condition arises from the need to address potential experiment-related effects within the artificial language learning paradigm. This concern is often addressed by employing two counterbalanced groups of participants, using different learning materials. By comparing the effects between these groups, researchers can attribute the experiment's impact to the manipulated variables rather than a preference for arbitrary unit combinations. In this study, the baseline condition serves to eliminate this alternative explanation, with the absence of significant rating differences across the three word types, suggesting that the design of the artificial language did not influence the experimental outcomes. If our design is correct, there will be no significant difference between these three types of words, which will initially be examined.

1.3 Does the length of exposure affect learning effects?

How to design the duration of the exposure phase is also an important issue in the SL filed. To ensure the detectability of the learning effect, many studies have employed extended

exposure phases in their experimental designs. For instance, Toro et al. (2005) repeated each nonsensical word 150 times, and Wang and Saffran (2014) repeated each nonsensical word 130 times in a tonal artificial language. Contradictory findings have emerged regarding the timing of the SL learning effect. Recent studies have challenged the assumption that long exposure times are requisite for SL to occur. These studies have used relatively shorter exposure phases and still observed pronounced SL effects (Qi et al., 2019; Arnon, 2020). For instance, adults and children as young as 7 to 9 years old displayed SL effects after only 32 repetitions, and older children aged 8 to 16 years old showed SL effects after 48 repetitions. In summary, the timing of when the learning effect occurs during the exposure phase remains an open question, and few studies have explored the trajectory of the learning effect observed in familiarity rating tasks in verbal SL field. The existing research landscape offers varying viewpoints on whether the learning effect emerges early in the exposure phase or requires extended exposure times. Further investigation is needed to clarify this aspect of SL processes, particularly in the context of verbal SL tasks.

1.4 The current study

The primary goal of the present study is to explore the independently explicit learning patterns associated with three distinct types of words and to examine how explicit knowledge exchange along with exposure time. To do so, the study manipulated three verbal statistical learning (SL) conditions:

1. Baseline Condition: The baseline condition served as a point of reference. It involved synthesizing syllables at random, without any occurrence of target words or partwords.

2. Short-Exposure Learning Condition (SEL): Within this condition, each word was repeated 45 times during the artificial language exposure.

3. Long-Exposure Learning Condition (LEL): In this condition, each nonsensical word was repeated a total of 90 times within the artificial language.

The study design employed a mixed-method approach, incorporating both within-subject (word type) and between-subject variables (learning condition). As an exploratory study, this study does not have a priori hypotheses. However, based on the discussions above, the planned tests are as follows: (1) examine the non-significant learning effect in the baseline condition, (2) compare the learning effects of the three word types between the SEL condition and the baseline condition to assess the explicit effect in verbal SL; and (3) compare the learning effects of the three word types between the SEL condition and the LEL condition to observe changes in the explicit effect over time.

2 Method

2.1 Participants

One hundred and forty-four native speakers of Mandarin (age range:18-28; females = 123) were first recruited from a University in Southeast China. Participants were randomly assigned to either the long-exposure learning condition (49 participants), the short-exposure learning condition (49 participants) or the baseline condition (46 participants). All participants were right-handed with no formal musical training and were not majoring in foreign languages. The experiments were approved by the Institutional Review Board of the institution, and all participants signed informed consents before starting the experiments.

2.2 Materials

In tone languages, research on statistical learning is still relatively scarce. The Mandarin phonological system includes a total of 413 syllables and 4 lexical tones (i.e., Tone 1, Tone 2, Tone 3, and Tone 4). They can then combine into about 1522 tonal syllables. For our experiment,

we followed the design of Gómez et al., (2017) and selected twelve syllables with Tone 1 to create nonsensical tonal syllables, thereby minimizing the influence of the different statistical information carried by tone and segmental structure. The syllables were recorded in a sound-attenuating room to digital format at 44100 Hz with 16bit precision. Twelve syllables were then normalized for duration (350ms), mean pitch (266Hz), and intensity (70dB) via Praat software. The 12 target syllables were randomly to form six disyllabic nonsensical words, as this type of words are most popular in Mandarin Chinese. Then, we created partwords and nonwords based on the target words respectively. The within-word, syllabic level TP for target words was 1.0, and the TP of syllables spanning across word boundaries was 0.2. Nonwords consisted of two syllables that never co-occurred during the exposure. The within-word TP for nonword syllables was therefore always 0. Three types of nonsensical words are shown in Table 1.

The artificial languages in the two learning conditions were created with the same pool of target words. In the LEL condition, six target words were utilized to create an artificial language stream, which resulted in 90 tokens of each target word and 18 token of partword. The same six words were used to create the artificial language in the SEL condition, which contained 45 tokens of each target word and nine tokens of partword. The LEL and SEL conditions were concatenated by a Praat script into a pseudorandom sequence, which ensured that the same word could not occur twice in a row. In the baseline condition, instead of consisting of six disyllabic words, the artificial language was concatenated with the same syllables that made up the other two conditions. The artificial languages lasted about six minutes for the LEL condition, but three minutes for the baseline and SEL condition. The test items in the three conditions were identical, with a total of 18 items across three types of words.

2.3 Procedure

All participants were told that they would hear an artificial language via headphones and would later be tested on their knowledge of the language. They then listened to the artificial language for either six or three minutes in a soundproof booth. After this exposure phase, a 6-point Likert scale familiarity rating task began (1 for not familiar at all and 6 for very familiar). Participants first took two practice trials, and then completed a total of 18 test trials. On each trial, participants were required to rate the familiarity of item considering the artificial language they had just listened to (see Fig.1). All three types of words (target words, partwords, and nonwords, with six of each type) appeared only once, and their presentation order in the trials was randomized across participants. No additional constraints were imposed to filter the participants' responses, and all trials were included in the analyses. The entire experiment was conducted using E-prime 3.0 and lasted approximately 10 minutes in all three conditions.

3 Results

3.1 Learning effect in baseline condition

We first investigated whether a learning effect existed among the three types of words under the baseline condition. To assess this, we used a Linear Mixed Model (LMM) using the lmer function in R (version 4.3.1), with word type as a fixed effect and both subject and item as random intercepts[1]. The ANOVA results showed no significant main effect of word type ($F(2,15) = 0.37$, $p = 0.70$), suggesting no substantial difference in learning among the three word types. Furthermore, the analysis revealed no significant fixed effects when comparing the familiarity ratings between target words and partwords ($\beta = -0.16$, $t = 0.62$, $p = 0.55$) or between partwords and nonwords ($\beta = -0.05$, $t = 0.19$, $p = 0.85$). These findings suggest that under the baseline condition, participants did not show a discernible learning effect or preference for any of the three word types based on their familiarity ratings.

3.2 SEL's learning effect compared with baseline condition

Another LMM model was established with condition and word type as fixed effect, item as a random slope, and both item and subject as random intercept[2]. One nonword in the SEL condition was incorrectly designed for 11 participants, resulting in only 17 trials for these participants, while others had 18 trials. The ANOVA results indicated a significant main effect of condition ($F(1,47.86) = 6.50$, $p = 0.014$) and a significant interaction effect between condition and word type ($F(2,15.00) = 17.53$, $p < 0.001$). The unstandardized coefficients of fixed effect are presented in Table 2. Next, we ran a series of post-hoc analysis with the function *emmeans*. All p values were Bonferroni adjusted when pairwise comparisons consisted of more than two levels. For target words, the rating scores in the SEL condition ($M = 4.57$) were significantly higher than those in the baseline condition ($M = 4.11$), $t = 2.29$, $\beta = 0.46$, $p = 0.03$. In contrast, partcipants rated partwords as more familiar in the baseline condition ($M = 4.27$) than in the SEL condition ($M = 3.49$), $t = 3.86$, $\beta = 0.78$, $p < 0.001$. A similar pattern was also found in the pairwise comparison of nonwords between the two SL conditions (baseline: $M = 4.32$, SEL condition: $M = 3.53$), baseliene – SEL condition: $t = 3.93$, $\beta = 0.79$, $p < 0.001$. See Fig.2 for a visualization of participants' rating patterns in the baseline and SEL conditions. These results suggest that participants began to demonstrate a learning effect at the beginning after the exposure phase, but the strength of this effect varied depending on the word type.

3.3 LEL's learning effect compared with SEL condition

We employed a similar method to examine the explicit learning effects of participants between SEL and LEL conditions. In the LMM model, word type and learning condition were treated as fixed effects, while subjects are set as random intercept, and item was set as both random intercept and random slope[3]. The results showed that only the main effect of word type

is significant ($F(2,15.03) = 17.63$, $p < 0.001$), while other main effect ($F(1,56.17) = 1.22$, $p = 0.28$) and interaction ($F(1,14.99) = 0.11$, $p = 0.89$) are not significant. Similarily, only two fixed effects reached significnace, partwords – target words: $t = –4.48$, $\beta = –1.08$, $p < 0.001$, nonword – target words: $t = –4.33$, $\beta = –1.04$, $p < 0.001$. The unstandardized Coefficients of Fixed effect could be seen in Table 3. See Fig.3 for a visualization of participants' rating patterns in the SEL and LEL condition.

3.4 Correlations across three types of words in three conditions

We further conducted a series of correlation analysis to demonstrate the learning effect under different conditions. In baseline condition, participants' rating scores on target words correlated significantly with that of partwords and nonwords (target words and partwords: $r = 0.33$, $p = 0.02$, target words and nonwords: $r = 0.64$, $p < 0.001$); meanwhile, the correlation coefficient between the latter two also reached a significant level (partwords and nonwords: $r = 0.49$, $p < 0.001$). This pattern of correlation matrix changed dramatically in SEL condition, specifically, only partwords and nonwords' rating scores correlated significantly with each other ($r = 0.67$, $p < 0.001$), but the other two correlations were not significant any more (target words and partwords: $r = 0.08$, $p = 0.59$, target words and nonwords: $r = 0.22$, $p = 0.13$). Finally, we replicated the above analysis in LEL condition and the results showed that target words' rating scores correlated significantly with partwords and nonwords' rating scores again (target words and partwords: $r = 0.43$, $p < 0.01$, target words and nonwords: $r = 0.46$, $p < 0.01$), as well as partwords' rating scores still showed a strong correlation relationship with nonwords' rating scores ($r = 0.71$, $p < 0.001$). The pattern of correlation relationship between three types of words across three learning conditions indicated that significant changes in learning effect have emerged from SEL condition to LEL condition, although the mean of rating scores did not

change. In these correlation analyses, all *p* values were Bonferroni adjusted when pairwise comparisons consisted of more than two levels. See Fig.4 for a visualization of correlation relationship between participants' rating in three learning condition.

4. Discussion

While previous research has employed the 2AFC task to establish explicit learning effects in verbal SL tasks across various participant groups, there has been a notable lack of studies exploring whether other tasks, such as familiarity rating tasks, could provide new insight into understanding the components of learning effects in statistical learning. The current study included a baseline condition to investigate the explicit learning effects of different types of words in verbal SL tasks. The findings of the study revealed an intriguing and contrasting pattern of learning effects: participants showed higher familiarity with target words and reduced familiarity with foils (both partwords and nonwords) in the short exposure condition compared to the baseline condition. In addition, this explicit effect did not change along with exposure time, indicating that the learning pattern was independent of the length of the exposure time.

4.1 The contrasting learning effect of targets versus foils

The initial assumption was that target words and partwords would receive higher familiarity ratings due to their repeated exposure during the learning phase, like the linear decline across three distinct word types as revealed in research of Batterink and Paller (2017), while nonwords would maintain consistent familiarity ratings across both learning and baseline conditions, given their absence in the exposure phase. However, the experimental results differed significantly from those expectations. Target words received significantly higher familiarity ratings in the short learning condition, whereas partwords and nonwords showed substantial decreases in familiarity ratings from the baseline condition to the SEL condition.

To our knowledge, previous studies have primarily focused on whether participants can recognize target words in 2AFC task, with little attention given to the learning effect of nonwords and partwords. As nonwords were deliberately excluded and partwords occurred infrequently during the learning phase of the artificial language, their lower familiarity ratings suggest that participants either recognized these items as not being stored in memory or noticed that they deviated from the statistical regularities of the artificial language. We propose that a type of self-awareness cognitive mechanism may have contributed to these results during the test phase. The artificial languages adopted in the current study are arguably easier than previous studies, as the TPs between target words is 0.2, which is lower than that in other studies (e.g., Batterink, 2017); the disyllabic nonsensical units also likely put less pressure on working memory, which is also an important factor in SL task difficulty (Palmer & Mattys, 2016). Given the relative simplicity of the task for adults and the instruction that learning would be assessed after the exposure phase, it is likely that participants consciously memorized the target words and ignored other words or constructions during exposure. Consequently, when partwords and nonwords were presented during the test phase without time constraint, participants were able to recall target words, and were therefore consciously aware that these items had not been learned. In other words, our results in the familiarity rating task demonstrated that these two types of foils underwent a similar cognitive process: a suppression effect from the baseline to the SEL condition. This self-awareness cognitive mechanism may be seen as a type of metacognition: the participants know what they have or have not learned, allowing them to explicitly reject the foils and resulting in lower familiarity ratings compared to the baseline condition. This explanation aligns with recent evidence showing that nonwords accepted by participants corresponded with lower confidence ratings, whereas rejected corresponded with higher confidence ratings

(Polyanskaya, 2022). The metacognition defined in this study is not exactly the same as the explicit mechanism proposed in previous studies. This is because the explicit instructions or supplemental training were given to the participants in previous studies (Batterink et al., 2015a; 2015b), so the SL's explicit mechanism runs through both the learning and the testing phase. In contrast, this study used implicit instructions: the knowledge of the language would be tested, so this kind of metacognition is more likely to appear at the testing stage.

Unlike the rating scores for foils, the rating scores of target words in the SEL condition reached approximately 4.8 on a scale of one to six, which is significantly higher than that in the baseline condition. This suggests that participants were able to successfully recognize the target words by tracking statistical regularities during the exposure phase. Combining the results of foils above, it is likely that the dual nature of target words' explicit memory leads to the contrasting rating patterns: recognition target words as well as rejection of partwords and nonwords.

In addition to the analysis based on group average, more attention should be given to the correlations between the three types of words across the baseline and SEL conditions. The high correlation between each pair of word types in the baseline condition indicates that the participants did not show a clear preference for any specific type but instead based their ratings on their familiarity with the syllables. This supports the validity of the baseline condition. In the SEL condition, the sharp changes in correlations between target words and partwords, as well as between target words and nonwords, suggest that participants consciously categorized the items into two groups: targets and foils. This resulted in high correlations within the same category and low correlations between different categories. In sum, our results demonstrate a distinction in familiarity ratings between target words and foils in verbal SL, and we propose that this finding

reflects a combination of explicit knowledge of the target words obtained in exposure phase and a form of metacognition in test phase.

Previous research has suggested that the learning effects observed in recall-based tasks like the 2AFC task reflect a composite of different abilities rather than isolating the targeted cognitive process (Frost et al., 2015; Isbilen et al., 2020), which is also susceptible to the influence of individual decision-making strategies (Christiansen, 2019; Isbilen & Christiansen, 2022). Because familiarity tasks allow participants to reflect what they have learned during the exposure phase, our findings provide direct evidence for this notion: when participants face a forced-choice trial between a partword and a target word, they may make decisions by eliminating foils rather than actively recognizing target words, or by employing both mechanisms. This contrasting pattern of familiarity ratings across the three types of words also adds depth to our understanding of the components underlying the learning effects traditionally measured by the 2AFC task. Based on the results of our study, the correctness of each trial in the 2AFC task can be further deconstructed into two parts: explicit memory of target words and the metacognitive knowledge about what has been learned. This suggests that the learning effects measured by the 2AFC task might overestimate the actual statistical knowledge gained during the exposure phase.

4.2 The change of learning effects along with exposure time

Another key aim of the present study was to determine whether an extended exposure phase could enhance the learning effect across the three types of words. Notably, the familiarity ratings for target words in the SEL condition, where each word was repeated 45 times during the exposure phase, were already significantly higher than those in the baseline condition. Intriguingly, when the exposure time was doubled in the LEL condition, with each word repeated 90 times, no additional learning effect was observed compared to the SEL condition. The rapid-

learning effect aligns with findings from other online studies. For instance, Batterink (2017) demonstrated that after just one exposure to words within continuous nonsensical speech, participants showed faster reaction times to final syllables compared to initial syllables. Similarly, a study by Siegleman and colleagues (2018) utilized a self-paced SL paradigm in the visual modality and highlighted the timing of SL effects, showing that in the visual domain, the learning effect followed a logarithmic function, with improved learning rates after as few as seven repetitions of each triplet. By situating these outcomes within the broader context, the present study also adds to the growing body of evidence supporting the idea of rapid learning in both visual and verbal SL tasks. The consistency between these findings further emphasizes the intriguing nature of rapid learning effects within the realm of SL processes.

Interestingly, the correlation between the rating scores of target words and those of partwords and nonwords returned to a significant level, similar to those in the baseline condition. Despite this, we did not think that they reflect the same cognitive mechanism. This outcome in LEL condition appears to be directly related to the reduced range of target word rating scores[4], as shown in the third row of Figure 3. The mean rating scores in the LEL condition remained unchanged, but participants rated target words with greater consistency, leading to a steeper data distribution. Although we suspect that the significance of the correlation relationship may be an artifact, its presence also hints at a distinct learning effect stemming from extended exposure time. In conclusion, these data reveal a learning effect not previously observed: extended learning time does not alter the average of familiarity ratings, but it does increase the consistency of those ratings. In the future, it would be valuable to investigate this unique effect. Key questions to explore include: When does this type of learning effect occur? and Is it an implicit knowledge or explicit knowledge?

5. Conclusions

Statistical learning is widely regarded as a learning process that incorporates both implicit and explicit mechanisms. The results of the familiarity rating task in this study provide valuable insights into the nature of the independent learning effect, which varies depending on the specific type of word, highlighting the influence of explicit knowledge obtained in exposure phase and metacognition in test phase. At the same time, we have discovered a new pattern of how SL effects vary with learning time. The study is the first to explore the explicit knowledge of SL with familiarity rating task and emphasizes the need to reassess the components of the learning effects as measured in previous studies.

Notes

1. We started the model trimming from a full model (i.e., stimulus item was set as both random slope and intercept), but we finally found that the model fit was only achieved under the condition where item was only set as random intercept. All models in the current study could be found in Rmd file on website of OSF.

2. The model was established with treatment contrast. Both target words and baseline condition were set as baseline level respectively in LMM. We constructed four models in total, among which only the model with subjects as random intercept and items as random slope and random intercept, as well as the simplest model without any random slopes, did not exhibit any singular values and were successfully fitted. Furthermore, the difference between these two models was not statistically significant ($\chi^2 = 3.99$, $p = 0.14$). In order to more accurately assess the experimental effects, we have decided to retain the model that includes random slopes as our final model.

3. The model was established with treatment contrast. Both target words and SEL condition were set as baseline level respectively in LMM. We constructed four models in total, among which only the model with participants as random intercept and items as random slopes and random intercept, as well as the simplest model without any random slopes, did not exhibit any singular values and were successfully fitted. Furthermore, the difference between these two models was not statistically significant (χ^2 = 4.32, p = 0.11). In order to more accurately assess the experimental effects, we have decided to retain the model that includes random slopes as our final model.

4. To compare the differences in the degree of variation in rating scores, we conducted a test for homogeneity of variance. For target words, the variance in LEL condition marginally significantly higher than that in SEL condition (F(1,96) = 2.86, p = 0.09). For partwords and nonwords, the results were not significant (partwords: F(1,96) = 1.38, p = 0.24, F(1,96) = 1.54, p = 0.22).

Acknowledgments

This work was supported by the Social Science Foundation of Jiangsu Province Higher Education Institutions [2022SJYB2051]; and the Initial Scientific Research Fund of Nanjing Normal University [184080H202A121].

Declaration of Interest Statement

The authors report there are no competing interests to declare.

Data Availability Statement

The data that support the findings of this study are openly available in OSF at [https://osf.io/xh6ju/?view_only=6f1659f166934a47b4f5494aa4025dd1]

Reference

Arnon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? an investigation of task reliability across modality. *Behavior Research Methods52* :68-81. https://doi.org/10.3758/s13428-019-01205-5

Batterink, L. J. (2017). Rapid statistical learning supporting word extraction from continuous speech. *Psychological Science, 28*(7), 921-928. https://doi.org/10.1177/0956797617698226

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, *90*, 31-45. https://doi.org/10.1016/j.cortex.2017.02.004

Batterink, L. J., Reber, P. J., & Paller, K. A. (2015a). Functional differences between statistical learning with and without explicit training. *Learning & Memory*, *22*(11), 544. https://doi.org/10.1101/lm.037986.114

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015b). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language, 83*, 62-78. https://doi.org/10.1016/j.jml.2015.04.004

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in cognitive science*, *11*(3), 468-481. https://doi.org/10.1111/tops.12332

Erickson, L. C., Kaschak, M. P., ED Thiessen, & Berry, C. (2016). Individual differences in statistical learning: conceptual and measurement issues. *Collabra*, *2*(1), 14. https://doi.org/10.1525/collabra.41

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in cognitive sciences, 19*(3), 117-125. https://doi.org/10.1016/j.tics.2014.12.010

Frost, R., Armstrong. B. C., & Christiansen. M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128-1153. https://doi.org/10.1037/bul0000210

Gómez DM, Mok, P., Ordin, M., Mehler, J., & Nespor, M. (2017). Statistical speech segmentation in tone languages: the role of lexical tones. *Language & Speech*, *61*(1), 84-96. https://doi.org/10.1177/0023830917706529

Isbilen, E. S., & Christiansen, M. H. (2022). Statistical Learning of Language: A Meta‐Analysis Into 25 Years of Research. *Cognitive Science, 46*(9), e13198. https://doi.org/10.1111/cogs.13198

Isbilen, E. S., McCauley, S. M., & Christiansen, M. H. (2022). Individual differences in artificial and natural language statistical learning. *Cognition, 225* (2022) 105123. https://doi.org/10.1016/j.cognition.2022.105123

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically induced chunking recall: a memory-based approach to statistical learning. *Cognitive Science*, 44(7). https://doi.org/10.1111/cogs.12848

Lukács, Á., Dobó, D., Szőllősi, Á., Németh, K., & Lukics, K. S. (2023). Reading fluency and statistical learning across modalities and domains: online and offline measures. *Plos one, 18*(3), e0281788. https://doi.org/10.1371/journal.pone.0281788

Mirman, D., Magnuson, J. S., Graf Estes, K., & Dixon, J. A. (2008). The link between statistical segmentation and word learning in adults. *Cognition*, *108*(1), 271-280. https://doi.org/10.1016/j.cognition.2008.02.003

Ordin, M., & Polyanskaya, L. (2021). The role of metacognition in recognition of the content of statistical learning. *Psychonomic Bulletin & Review, 28*, 333-340. https://doi.org/10.3758/s13423-020-01800-0

Polyanskaya, L. (2022). Cognitive mechanisms of statistical learning and segmentation of continuous sensory input. *Memory & Cognition*, *50*(5), 979-996. https://doi.org/10.3758/s13421-021-01264-0

Palmer, S. D., & Mattys, S. L. (2016). Speech segmentation by statistical learning is supported by domain-general processes within working memory. *The Quarterly Journal of Experimental Psychology, 69*(12), 2390-2401. https://doi.org/10.1080/17470218.2015.1112825

Qi, Z., Sanchez Araujo, Y., Georgan, W. C., Gabrieli, J. D., & Arciuli, J. (2019). Hearing matters more than seeing: A cross-modality study of statistical learning and reading ability. *Scientific Studies of Reading, 23*(1), 101-115. https://doi.org/10.1080/10888438.2018.1485680

Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality‐based differences in the effect of age. *Developmental Science*, *21*(4), e12593. https://doi.org/10.1111/desc.12593

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive science*, *37*(7), 1290-1320. https://doi.org/10.1111/cogs.12050

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928. https://doi.org/10.1126/science.274.5294.1926

Shoaib, A., Wang, T., Hay, J. F., & Lany, J. (2018). Do infants learn words from statistics? evidence from English‐learning infants hearing Italian. *Cognitive Science*, *42*(8), 3083–3099. https://doi.org/10.1111/cogs.12673

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: current pitfalls and possible solutions. *Behaviour Research Methods, 49*(2), 1-15. https://doi.org/10.3758/s13428-016-0719-z

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018). Redefining "learning" in statistical learning: what does an online measure reveal about the assimilation of visual regularities? *Cognitive Science, 42*(3), 692-727. https://doi.org/10.1111/cogs.12556

Toro, J. M., Pons, F., Bion, R. A. H., & Sebastián-Gallés, N. (2011). The Contribution of Language-Specific Knowledge in the Selection of Statistically-Coherent Word Candidates. *Journal of Memory and Language, 64*(2), 171–180. https://doi.org/10.1016/j.jml.2010.11.005

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97*(2), B25-B34. https://doi.org/10.1016/j.cognition.2005.01.006

von Koss Torkildsen, J., Arciuli, J & Ona Bø Wie. (2019). Individual differences in statistical learning predict children's reading ability in a semi-transparent orthography. *Learning and Individual Differences*, *69*(2019), 60-68. https://doi.org/10.1016/j.lindif.2018.11.003

Wang, T. L., & Saffran, J. R. (2014). Statistical learning of a tonal language: the influence of bilingualism and previous linguistic experience. *Frontiers in Psychology*, 5. https://doi.org/10.3389/fpsyg.2014.00953

Table 1 test items in three SL conditions

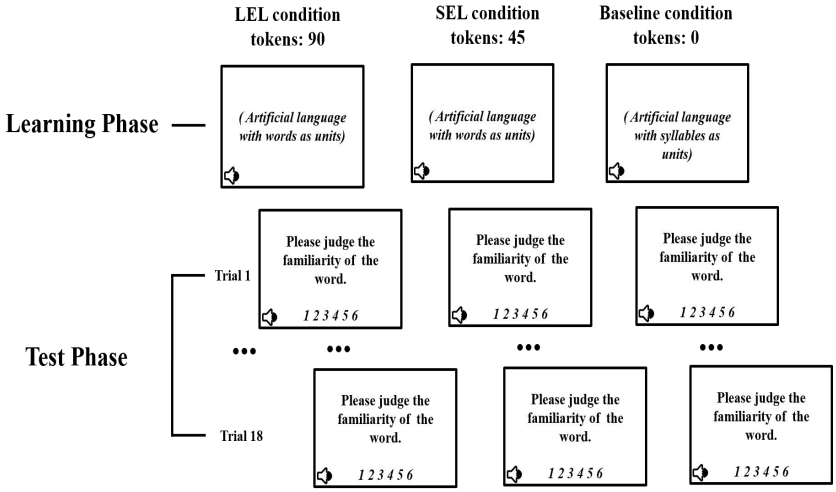
target word	partword	nonword
mei1nei1	se1mei1	rao1re1
[meɪ1neɪ1]	[sə1meɪ1]	[raʊ1re1]
rao1dia1	dia1re1	ruo1lai1
[raʊ1tɪa1]	[tɪa1re1]	[rɒŋ1laɪ1]
ruo1se1	nei1te1	mei1te1
[rɒŋ1se1]	[neɪ1t'e1]	[meɪ1t'e1]
lai1fo1	nue1ruo1	se1nei1
[laɪ1fo1]	[nyɛ1ruo1]	[sen1eɪ1]
te1nue1	fo1rao1	nue1rou1
[t'e1nyɛ1]	[fo1raʊ1]	[nyɛ1ruo1]
re1rou1	rou1lai1	fo1dia1
[re1ruo1]	[ruo1laɪ1]	[fo1tɪa1]

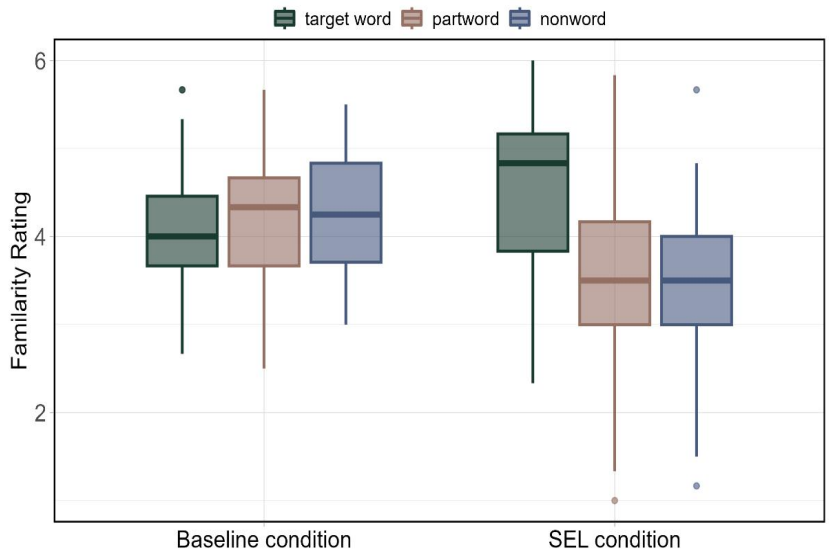
Table 2. The unstandardized Coefficients of Fixed effect in LMM model (*estimate*, *SE*, *t* value, and *p* value)

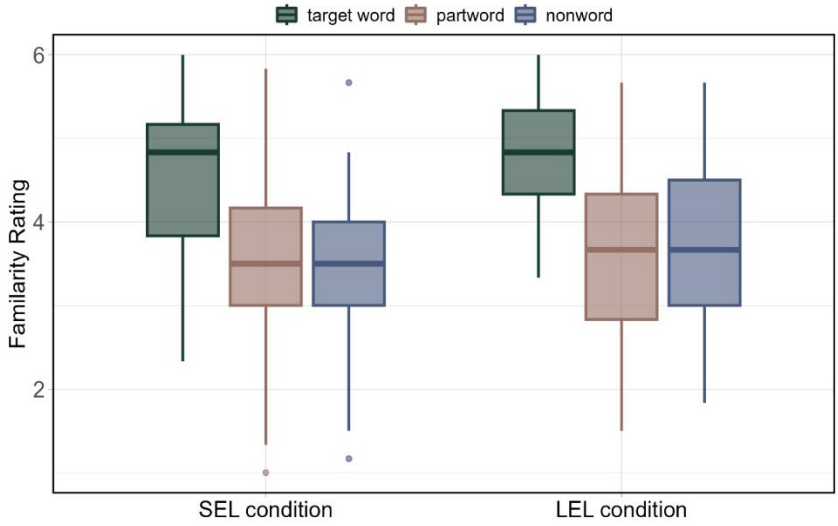
Fixed effect	*Estimate*	*SE*	*t*	*p*
Intercept	4.10	0.20	21.01	< 0.001
Condition SEL condition	0.46	0.20	2.29	0.03
word type: partword	0.16	0.26	0.63	0.54
word type: nonword	0.21	0.26	0.82	0.42
Condition SEL condition: word type partword	-1.24	0.24	-5.10	< 0.001
Condition SEL condition: word type nonword	-1.25	0.24	-5.16	< 0.001

Table 3. The unstandardized Coefficients of Fixed effect in LMM model (*estimate*, *SE*, *t* value, and *p* value)

Fixed effect	*Estimate*	*SE*	*t*	*p*
Intercept	4.57	0.19	24.03	< 0.001
Condition SEL condition	0.24	0.21	1.12	0.27
word type: partword	-1.08	0.24	-4.48	<0.001
word type: nonword	-1.04	0.24	-4.33	<0.001
Condition SEL condition: word type partword	-0.12	0.25	-0.47	0.64
Condition SEL condition: word type nonword	-0.07	0.25	-0.30	0.77


LEL condition
tokens: 90
SEL condition
tokens: 45
Baseline condition
tokens: 0
Learning Phase
(Artificial language with words as units)
(Artificial language with words as units)
(Artificial language with syllables as units)
Test Phase
Trial 1
Trial 18
Please judge the familiarity of the word.
1 2 3 4 5 6


target word
partword
nonword
Familarity Rating
6
4
2
Baseline condition
SEL condition


target word
partword
nonword
Familarity Rating
6
4
2
SEL condition
LEL condition

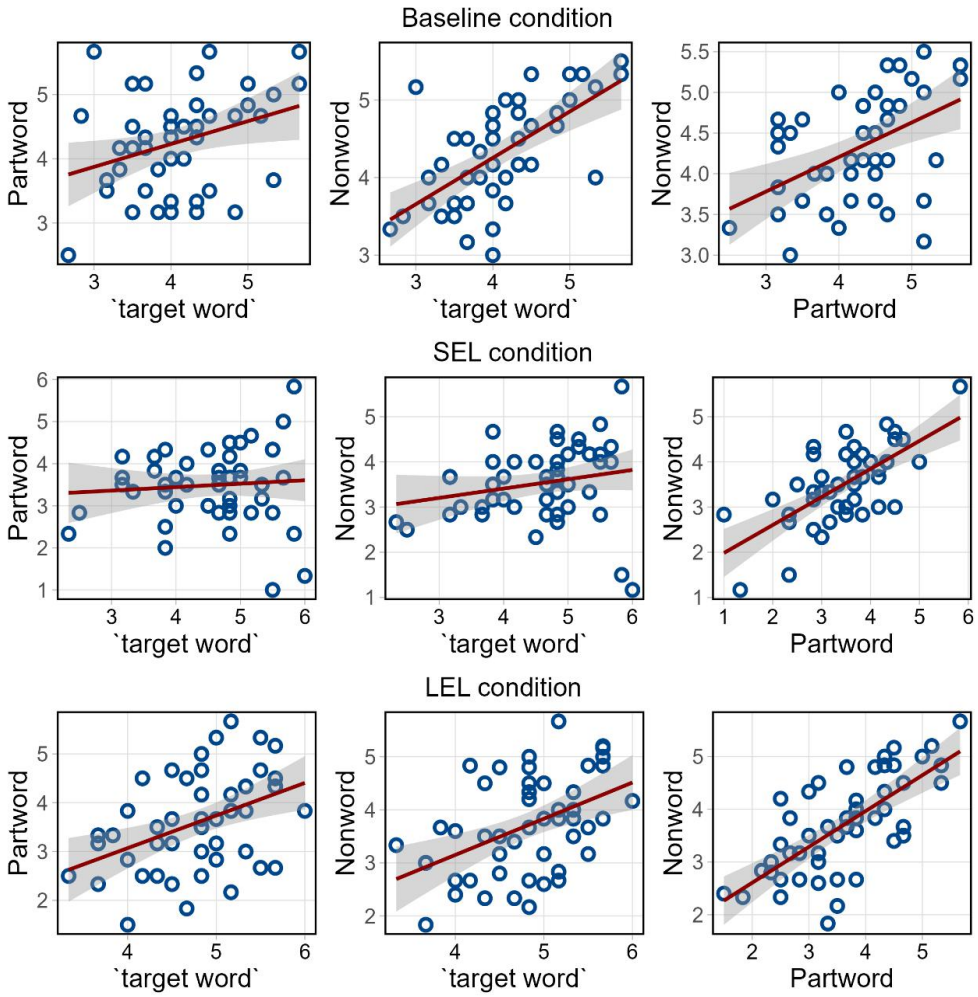


Figure	Figure Captions
Fig.1	Schematic representation of three conditions of verbal SL task
Fig.2	Familiarity ratings across word types in baseline and SEL conditions
Fig.3	Familiarity ratings across word types in SEL and LEL conditions
Fig.4	Correlations of familiarity ratings across word types in three conditions